Exploiting KonText for querying Lindat corpora

Natalia Klyueva

Institute of Formal and Applied Linguistics
Faculty of Mathematics and Physics
Charles University in Prague

kljueva@ufal.mff.cuni.cz

February 22, 2016
Overview

1 Introduction
   - Lindat repository
   - KonText

2 Querying Lindat corpora
   - Search in Universal Dependencies
   - Search in PDT
   - Other corpora

3 Conclusion
1 Introduction
   • Lindat repository
     • KonText

2 Querying Lindat corpora
   • Search in Universal Dependencies
   • Search in PDT
   • Other corpora

3 Conclusion
Ludwig van Beethoven, 1801

There ought to be only one grand dépôt of art in the world, to which the artist might repair with his works, and on presenting them receive what he required...

<table>
<thead>
<tr>
<th>Year</th>
<th>Views</th>
<th>Downloads</th>
</tr>
</thead>
<tbody>
<tr>
<td>2015</td>
<td>111021</td>
<td>49471</td>
</tr>
<tr>
<td>2016</td>
<td>40457</td>
<td>3698</td>
</tr>
</tbody>
</table>
Repository items

corpora

LINDAT

lexical resources

tools, services
1 Introduction
   - Lindat repository
   - KonText

2 Querying Lindat corpora
   - Search in Universal Dependencies
   - Search in PDT
   - Other corpora

3 Conclusion
- Corpus query tool
- first UI - Bonito
- SketchEngine: commercial and NoSke
- KonText - developed by the Institute of Czech National Corpus, based on SketchEngine
- Lindat-KonText: syntactic annotation and tree view
Je tu pro vás připravena rubrika Daňový poradce.
With President Karzai certain to win the elections and the demotion of key former Northern Alliance figures such as General Fahim and warlord Ismail Khan, there is now little reason for moderate Taliban leaders to fear reprisals from former Northern Alliance figures if they return home.
- Corpus Query Language [attribute="value"]
- Traditionally: form lemma tag + some derived attributes [tag="Vp.*"] [lemma="matka"]
- more complex annotation in Lindat Treebanks [tag="Vp.*"] [p_afun="Coord" & tag="NNF.*"]
- more information on Search in KonText see the Czech National Corpus wiki
More attributes for syntax

- word
- lemma
- tag
- afun (deprel)

- p_form
- p_lemma
- p_tag
- p_afun
  parent="+3"

- ep_form
- ep_lemma
- ep_tag
- ep_afun
  eparent="-5"
1 Introduction
   - Lindat repository
   - KonText

2 Querying Lindat corpora
   - Search in Universal Dependencies
   - Search in PDT
   - Other corpora

3 Conclusion
Universal Dependencies 1.2 - treebanks in 38 languages

- Google universal part-of-speech tags: [pos="ADJ"]
- Interset, universal features (ufeat). Ex., they: Case=Nom|Number=Plur|Person=3|PronType=Prs
- Stanford dependencies: deprel="root"

Lindat-KonText
ADJ: adjective
ADP: adposition
ADV: adverb
AUX: auxiliary verb
CONJ: coordinating conjunction
DET: determiner
INTJ: interjection
NOUN: noun
NUM: numeral
PART: particle
PRON: pronoun
PROPN: proper noun
PUNCT: punctuation
SCONJ: subordinating conjunction
SYM: symbol
VERB: verb
X: other
- Animacy: animacy
- Aspect: aspect
- Case: case
- Definite: definiteness or state
- Degree: degree of comparison
- Gender: gender
- Mood: mood
- Negative: whether the word can be or is negated
- NumType: numeral type
- Number: number
- Person: person
- Poss: possessive
- PronType: pronominal type
- Reflex: reflexive
- Tense: tense
- VerbForm: form of verb or deverbal
- Voice: voice
acl: clausal modifier of noun (adjectival clause)
advcl: adverbial clause modifier
advmod: adverbial modifier
amod: adjectival modifier
appos: appositional modifier
aux: auxiliary
auxpass: passive auxiliary
cc: coordinating conjunction
compound: compound
conj: conjunct
cop: copula
csubj: clausal subject
csubjpass: clausal passive subject
det: determiner
discourse: discourse element ... and many more
The examples here are just illustration of queries. It is not a meaningful linguistic research!!!
Examples

- LOGIN -> Shibboleth
- Position of adjectives in the Romance languages
  \[
  \text{[pos="ADJ" \& p\_pos="NOUN" \& parent="\textbackslash +.*"]}
  \]
- In English, a predicate before a subject
  \[
  \text{[deprel="nsubj" \& p\_deprel="root" \& p\_pos="VERB" \& p\_lemma!="be" \& parent="-.*"]}
  \]
1 Introduction
   - Lindat repository
   - KonText

2 Querying Lindat corpora
   - Search in Universal Dependencies
   - Search in PDT
   - Other corpora

3 Conclusion
- morphological layer
- analytical layer
- tectogrammatical layer
PDT attributes

- **Node attributes:**
  - m-layer, m-layer: word, lemma, tag
  - a-layer: ord, clause_number, is_member, afun, a_type
  - t-layer: deepord, t_lemma, functor, tfa, sempos, grammatemes_rf, coref_special, antes, discourse_special, discourse_type, discourse_target
  - parent: p_form, p_lemma, p_tag, p_afun, parent
  - eparent: ep_form, ep_lemma, p_tag, ep_afun, eparent
  - more see the PDT manual
Word order in Czech

- **SV**
  
  \[afun="Sb" \& p\_afun="Pred" \& parent="\+.*"]

- **VS**
  
  \[afun="Sb" \& p\_afun="Pred" \& parent="\-.*"]

- "se" stands more than 10 positions from the verb:
  
  \[lemma="se" \& ep\_afun="Pred" \& parent="\+(1|2).+"]

- want more - e.g. SVO? Better to use PMLTQ...

Auxiliary word, member of coordination:

\[is\_member="1" \& a\_type="aux"]
functors:
[tag="N...[^7].*" & functor="MEANS"]
[functor="DPHR"]{3}

tfa:
[afun="Sb" & tfa="f" & p_afun="Pred" & parent="\-.*"]

coreference:
[coref_special!="_"]

discourse:
[discourse_type="opp"]
searching in trees with CQL
Querying other corpora

- Czech Legal Text Treebank
  - lemma="se" & ep_tag="V.*"
- speech corpora:
- corpus with sentiment analysis
  - word=":)") within <s polarity="n" />
1 Introduction
   • Lindat repository
   • KonText

2 Querying Lindat corpora
   • Search in Universal Dependencies
   • Search in PDT
   • Other corpora

3 Conclusion
Conclusion and plans for the nearest future

- KonText adjusted to search in Lindat corpora
- Added tree view functionality
- Plans for nearest future:
  - make all the corpora from Lindat available via KonText
  - parse the corpora without annotation
  - PDT attributes from t-layer
  - PDT - tree view for a-trees and t-trees
  - add a short manual for each corpus
- Long-term goals:
  - word sketches
  - connection to the dictionaries, valency lexicon
  - word alignments for parallel corpora
Thank You!